Title
Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian

Journal
Water Resources Research, 39(5)

ISSN
0043-1397

Authors
Beven, K
Young, P
Gupta, H
et al.

Publication Date
2003

DOI
10.1029/2001WR001183

Comment on "Bayesian recursive parameter estimation for hydrologic models" by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian

Keith Beven and Peter Young

Environmental Science, Lancaster University, Lancaster, UK

Received 16 January 2002; accepted 1 July 2002; published 7 May 2003.

INDEX TERMS: 1860 Hydrology: Runoff and streamflow; 1894 Hydrology: Instruments and techniques; 9820 General or Miscellaneous: Techniques applicable in three or more fields; *KEYWORDS:* recursive estimation, uncertainty, rainfall-flow models, GLUE, model errors

Citation: Beven, K., and P. Young, Comment on ''Bayesian recursive parameter estimation for hydrologic models'' by M. Thiemann, M. Trosset, H. Gupta, and S. Sorooshian, *Water Resour. Res.*, *39*(5), 1116, doi:10.1029/2001WR001183, 2003.

[1] We believe that the paper by *Thiemann et al.* [2001] (hereinafter referred to as TTGS), while formally correct within the limitations of the underlying assumptions of their analysis, does not lead to sensible results in respect of the uncertainties normally associated with the parameters of hydrological models. Moreover, we feel that where these assumptions are reasonable, there may be more effective ways of achieving the same ends. In effect, the BARE methodology and the GLUE approach that they contrast it with, represent two extreme positions in a spectrum of possible Bayesian approaches to the problem. For instance, the results presented in the paper suggest that, in the BARE algorithm as implemented by TTGS, all of the error is treated as if it were ''measurement error'' and the estimated model parameters are obtained without any appreciable uncertainty. In other words, the estimated model appears, from the TTGS results, to be an almost deterministic, or true, representation of the system. In GLUE, on the other hand, the sources of error are treated implicitly in weighting the predictions of multiple (non-error free) models, without strong assumptions about a measurement error model (which might indeed vary from model realization to model realization). However, both of these extremes are only partially adequate: a preferable approach would be to separate out the effects of errors in the inputs, the errors in the model structures and real measurement errors in the outputs.

[2] The most critical issue raised by the TTGS paper is the implicit assumption that the model structure is correct. In particular, the BARE parameter estimates, as reported by TTGS, effectively converge to point estimates with apparently little parameter uncertainty. This result is surely difficult to justify in any practical terms; indeed, it certainly cannot be justified in the first simulation example (which involves a linear, two term Nash Cascade (NC) model with assumed known rainfall input and white, Gaussian measurement noise). In this case, the presence of additive noise means that the parameter estimates should converge to a normal probability distribution with mean and covariance that depend on the number of samples that have been processed. Moreover, the required parameter estimates and their associated covariance matrix can be estimated in a much more computationally efficient manner than in BARE, using a recursive Bayesian parameter estimation algorithm in which only the first two moments of the distribution (mean and covariance) are updated sequentially, so that no Monte Carlo simulation is necessary at all.

[3] For example, the optimal, refined instrumental variable (RIV) parameter estimation algorithm [*Young*, 1984] yields recursively updated Maximum Likelihood (ML) estimates of the parameters in an unconstrained transfer function model [*Young*, 2002a]. In this example, the NC is a second order transfer function that is constrained (the two first order elements are identical), so the standard RIV estimates are sub-optimal in an ML sense. Even so, the algorithm clearly identifies that the TF is second order (using the YIC criterion [*Young*, 1990]) and the recursive estimates of the TF parameters (true values: 0.8521 and 0.0219) based on the estimates obtained at the end of the data set are 0.8568(0.0124) and 0.0214(0.0009). Here the figures in brackets are the estimated standard errors and the estimate of the eigenvalue at 0.8521, is based on the mean of the two estimated eigenvalues (normally slightly different because the estimates of the TF parameters are unconstrained). If complete optimality is required, then it is necessary to use a more complex, constrained ML numerical optimization procedure (still much simpler than BARE) and this yields estimates 0.8511(0.0023) and 0.0213(0.0007). Converted to the residence time (or recession coefficient) estimate, as obtained by BARE (true value 25), the unconstrained RIV estimates at the end of the data set are 25.8766(+2.6574, −2.2340) and the ML estimates are 24.8187(+0.4221, −0.4093). (The range is given here because the transformation of the eigenvalue estimate to the residence time estimated distorts the error distribution.) Clearly, uncertainty in the model parameters still exists at the end of the data set and this is what would be expected for any finite set of data with measurement noise. Finally, in contrast to the BARE situation shown in figure 3 of the TTGS paper, the estimated standard error band on the model predicted output encompasses all of the modeling errors, as it should do [see *Young*, 2002a].

[4] It is worth noting that, in practical terms, the complication of constrained estimation in the NC example would not be necessary, except in exceptional circumstances, because the linear, identical element, NC is not normally considered a good TF model of the obviously nonlinear

0043-1397/03/2001WR001183

rainfall-flow process. A more appropriate model is a nonlinear TF model in which nonidentical TF elements are connected in parallel, rather than serial form [e.g., *Jakeman et al.*, 1990; *Young*, 2001a, and references therein]. In this case, the RIV algorithm is optimal for white or colored, Gaussian measurement errors; and consistent (asymptotically unbiased) for other non-Gaussian noise processes (see later comments). For example, using the same rainfall data used in the NC simulation example but modeling in terms of the actual measured flow, rather than the NC simulated flow [*Young*, 2002a], results in a third order nonlinear model of the parallel flow type mentioned above, with a coefficient of determination (Nash-Sutcliffe efficiency) of $R_T^2 = 0.865$; whereas for the equivalent third order NC model, the explanation of the data is poor, with $R_T^2 = 0.647$.

[5] The convergence of the model parameter values to (essentially) a single parameter set is repeated for the real world example, where the SAC-SMA model is applied to the Leaf River catchment, as shown in Figure 6 of TTGS This is certainly a convenient outcome for all sorts of reasons: for example, it is possible to use the set of parameter values to relate to catchment characteristics in some regionalization exercise; or to change one or more parameters to reflect some aspect of change in the catchment; or simply to compare different sets of deterministic outputs under different input conditions, such as might result from climate change scenarios. However, we do not see how such an outcome can be a justified conclusion in this case. There are (unknown and time variable) errors in the inputs to the model; there are (unknown and probably time variable) errors in the model structure; and there are (unknown and time variable) errors in the observed model outputs. The result (particularly under the additionally restrictive assumption that the errors are independent) is a dramatic overconditioning of the model to a single parameter set.

[6] It is well known that ignoring correlation in the errors (as is only too evident in Figure 5 of TTGS) should be expected to lead to biased parameter estimates unless the model includes a stochastic representation of the correlation in the errors (i.e. a noise model is included); or, more simply and robustly, an instrumental variable estimator is used (see above). There is a strong hint of this overconditioning in Figure 6 of TTGS, which reveals significant jumps in the high probability density region of individual parameter values, prior to settling down to their final values (as noted by the authors in their final discussion). In effect, it would appear that small changes in the calibration data (or possibly even in the input and output measurement errors) would result in convergence to quite a different set of parameter values. In addition, the final model (plus error estimates) does not bracket the observed discharge data for parts of nearly all the hydrographs in the calibration period; again an indication that their assumptions in this regard may not be justified. In addition, the authors give no demonstration that the parameters would converge to the same parameter set in another calibration period, which would be implied if there is no estimated uncertainty on the parameters.

[7] It is also well known that most of the difficulties associated with the calibration problem for rainfall-runoff models comes from assuming conceptual model structures a priori that are not reasonable representations of the system and are effectively overparameterized with respect to the information content of the (non-error free) input and output data series. This is still a common practice, despite the problems of trying to find global optimal parameter sets with either single or multiple objective functions (as amply demonstrated by previous work of the Arizona group). There are several responses to these problems.

[8] The first is to improve the identifiability of the chosen model. This is a viable strategy where only discharge prediction is required. The data-based mechanistic (DBM) methodology of Young [e.g., *Young*, 1998, and references therein], in which a parsimonious model is determined from the nature of the data itself, will lead to significantly improved identifiability in many cases. Applications to the rainfall-runoff problem are given by *Young* [2001a, 2001b, 2002b]: see also the nonlinear TF modeling results discussed above in relation to NC simulation model example. As in the NC example, such parsimonious models are easily set within a recursive estimation framework. Although, in the standard application of such an approach, this yields only the first two moments of the distribution, it can yield reasonable results in practical situations at a very modest computational cost when compared with the alternative, computationally expensive BARE approach, with its requirement for extensive Monte Carlo simulation.

[9] The second is to improve identifiability by the use of ''better'' single and multiple performance measures. The Arizona group pursued this strategy with the use of the Pareto optimal set approach of *Yapo et al.* [1998] and *Gupta et al.* [1998]. We would agree that the commonly used Nash-Sutcliffe efficiency criterion is not an adequate measure, despite its common usage, for many reasons. These include the effects of timing errors, the possibility of obtaining biased parameter estimates, and the possibility that it underestimates the information content of the data. However, we would also argue that the formal Bayesian measure used in this paper appears to grossly overestimate the information content of the data by underestimating the effects of the noise, resulting in the overconditioning evident in the results.

[10] The third strategy is to confront the problem of model structural error directly. It is true that this can be incorporated into the formal Bayesian framework as demonstrated by *Kennedy and O'Hagan* [2001] who incorporate a ''model inadequacy'' function into the identification process (although the most complex inadequacy function they include is a constant bias!). A full acceptance of model structural error, however, requires a change in philosophy because it means that the calibration problem can no longer be considered as simply a matter of finding the parameter values of the model [*Beven*, 2001c, 2001d]. This is the philosophy that underlies the GLUE methodology. Rather than being a rather poor approximation to the formal Bayesian methodology as suggested by TTGS, GLUE is (as the name suggests) intended as a wider-ranging methodology that accepts the possibility of equifinality of models, and allows for multiple competing model structures and parameter sets. From this perspective, the Bayesian methodology proposed in the TTGS paper is a special case of GLUE, within which the formal assumptions about the model structure (that it is true) and error structure (additive, independent and converging to near constant (transformed)

variance) can be justified and tested. The results presented by TTGS suggest that in the application to the Leaf River example, the formal Bayesian methodology cannot be justified, despite its apparent objectivity.

[11] The question then is what is an appropriate likelihood measure for use in a Bayesian context. GLUE is often criticized (as by TTGS) for being subjective in this respect and not properly Bayesian. However, just as the use of subjective priors is perfectly acceptable, the use of subjective likelihood measures does not preclude a Bayesian analysis [see *Howson and Urbach*, 1993]. It just means that the posterior cannot be considered a "true" probability distribution. It is rather a representation of subjective probabilities or probabilities that are conditional on the specific (and explicit) set of assumptions made [*Beven et al.*, 2000; *Beven*, 2000, 2001a, 2001b, 2001c, 2002a; *Beven and Freer*, 2001]. As well demonstrated in this paper, being objective about the likelihood measure does not mean that true probabilities result if the assumptions are incorrect.

[12] The differences, however, perhaps run a little deeper and here we disagree, even between ourselves, about what might be an appropriate strategy. In formal Bayes theory, as used by TTGS(2001), the posterior likelihood is intended to represent the probability of predicting an observation, given the (true) model, $L(Y|\Theta)$ where Y is the observation vector and Θ is a parameter vector. However, Bayes equation can be stated in the more general conditioning form for hypotheses H given evidence E as

$$L_p(H|E) \propto Lo(H)L(E|H)$$

or in the discrete form for k potential hypotheses proposed independently by Laplace as:

$$L_p(H_k|E) \propto L_o(H_k)L(E|H_k)/\Sigma_k L(H_k|E)$$

Where $L_p(H_k|E)$ is the posterior likelihood for hypothesis H_k given the evidence E; $L_o(H_k)$ is a prior likelihood for H_k; and $L(E|H_k)$ is the likelihood of predicting the evidence E given H_k [see *Howson and Urbach*, 1993; *Bernardo and Smith*, 1994]. Allowing that feasible models can be treated as hypotheses in this sense, then the GLUE methodology aims to estimate the likelihood of different models reproducing the observables, rather than the likelihood of an error given some true model (i.e., $L_p(H_k|E)$ as $L(\Theta_k|Y)$, rather than as $L(Y|\Theta)$). It can equally be applied on a recursive time step by time step basis (although it has usually been applied recursively over a block of time steps or calibration periods rather than single time steps).

[13] Thus the predictive uncertainty estimated by GLUE can be considered to result from the uncertainty of models at the expense of estimating errors associated with individual models. Errors resulting from both input and output measurements are then treated implicitly in the estimation of the likelihood associated with a model. Estimating a measure of belief in a model is really the much more interesting problem. As hydrologists, we are, after all, really much more interested in identifying the set of model representations that are consistent with our observations, than we are in soaking up the prediction errors in some (nonphysical) statistical error model [*Beven*, 2002b]. Different model structures as well as different parameter sets can be included within this framework, subject only to computationally constraints. Different error structures could actually be included as well, where this might have some advantage, for example in improving short term forecasts [see *Romanowicz et al.*, 1994, 1996], but at the expense of introducing additional parameters of the error model and computational complexity.

[14] In their response to this comment, TTGS note that the BARE approach is not limited to cases that assume the model is correct, and that the convergence to a single parameter set is largely a result of the limited sampling used in the current implementation and should not be interpreted as suggesting that there is no residual uncertainty, only that the final model was identified as the most likely of those available. Formally, this may be true and it may be that the essence of the argument is what likelihood function/measure(s) to use in the face of both data errors and model structural errors. We are still worried that a Bayesian estimation algorithm, applied to data that are admittedly noisy and using a model with undoubted structural error, should result in estimates with little or no uncertainty. Surely this is unacceptable in any stochastic analysis?

[15] We cannot currently separate all the potential sources of error without making very strong and specific assumptions about their nature. In BARE we would argue that the assumptions are too strong and cannot be justified; in GLUE it is not entirely satisfactory that they are treated implicitly in the model evaluation. It is perhaps in the middle ground between these two approaches, taking full advantage of modern techniques for recursively identifying and estimating parametrically parsimonious models, probably combined with Monte Carlo-based estimation (e.g., particle-filter methods and the estimation procedures suggested by *Kitagawa* [1996]), where future advances in calibration methodologies might be found. We would join with TTGS in hoping to stimulate a new generation of graduate students to start to think deeply about the problems of how to identify models of appropriate complexity in applications to difficult environmental systems.

[16] **Acknowledgments.** The authors are most grateful to Hoshin Gupta for providing the relevant Leaf River data set.

References

Bernardo, J. M., and A. F. M. Smith, *Bayesian Theory*, John Wiley, New York, 1994.

Beven, K. J., Uniqueness of place and process representations in hydrological modelling, *Hydrol. Earth Syst. Sci.*, *4*(2), 203–213, 2000.

Beven, K. J., How far can we go in distributed hydrological modelling?, *Hydrol. Earth Syst. Sci.*, *5*(1), 1–12, 2001a.

Beven, K. J., Calibration, validation and equifinality in hydrological modelling, in *Model Validation: Perspectives in Hydrological Science*, edited by M. G. Anderson and P. D. Bates, pp. 43–55, John Wiley, New York, 2001b.

Beven, K. J., Comment on M Kennedy and A O'Hagan, Bayesian calibration of computer models, *J. R. Stat. Soc., Ser. B*, *63*, 456, 2001c.

Beven, K. J., *Rainfall-Runoff Modelling: The Primer*, John Wiley, New York, 2001d.

Beven, K. J., Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system, *Hydrol. Processes*, *16*, 189–206, 2002a.

Beven, K. J., Towards a coherent philosophy for environmental modelling, *Proc. R. Soc. London, Ser. A.*, *458*, 2465–2484, 2002b.

Beven, K. J., and J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems, *J. Hydrol.*, *249*, 11–29, 2001.

Beven, K. J., J. Freer, B. Hankin, and K. Schulz, The use of generalised likelihood measures for uncertainty estimation in high order models of environmental systems, in *Nonlinear and Nonstationary Signal Processing*, edited by W. J. Fitzgerald et al., pp. 115–151, Cambridge Univ. Press, New York, 2000.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, Toward improved calibration of hydrologic models: Multiple and incommensurable measures of information, *Water Resour. Res.*, *34*, 751–763, 1998.

Howson, C., and P. Urbach, *Scientific Reasoning: The Bayesian Approach*, 2nd ed., 470 pp., Open Court, Chicago, Ill., 1993.

Jakeman, A. J., I. G. Littlewood, and P. G. Whitehead, Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments, *J. Hydrol.*, *117*, 275–300, 1990.

Kennedy, M. C., and A. O'Hagan, Bayesian calibration of mathematical models, *J. R. Stat. Soc., Ser. B*, *63*, 425–450, 2001.

Kitagawa, G., Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *J. Comput. Graphical Stat.*, *5*, 1–25, 1996.

Romanowicz, R., K. J. Beven, and J. Tawn, Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach, in *Statistics for the Environment II: Water Related Issues*, edited by V. Barnett and K. F. Turkman, pp. 297–317, John Wiley, New York, 1994.

Romanowicz, R., K. J. Beven, and J. Tawn, Bayesian calibration of flood inundation models, in *Floodplain Processes*, edited by M. G. Anderson, D. E. Walling, and P. D. Bates, pp. 333–360, John Wiley, New York, 1996.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian, Bayesian recursive parameter estimation for hydrologic models, *Water Resour. Res.*, *37*, 2521–2536, 2001.

Yapo, P. O., H. Gupta, and S. Sorooshian, Multi-objective global optimisation for hydrologic models, *J. Hydrol.*, *204*, 83–97, 1998.

Young, P. C., *Recursive Estimation and Time Series Analysis*, Springer-Verlag, New York, 1984.

Young, P. C., Recursive estimation, forecasting and adaptive control, *Control Dyn. Syst. Adv. Theory Appl.*, *30*, 119–166, 1990.

Young, P. C., Data-based mechanistic modelling of environmental, ecological, economic and engineering systems, *Environ. Modell. Software*, *13*, 105–122, 1998.

Young, P. C., Data-based mechanistic modelling and validation of rainfall-flow processes, in *Model Validation: Perspectives in Hydrological Science*, edited by M. G. Anderson and P. D. Bates, pp. 117–161, John Wiley, New York, 2001a.

Young, P. C., The identification and estimation of nonlinear stochastic systems, in *Nonlinear Dynamics and Statistics*, edited by A. I. Mees, pp. 127–166, Birkhäuser Boston, Cambridge, Mass., 2001b.

Young, P. C., Modelling the Leaf River data, *Rep. 177*, Cent. for Res. on Environ. Syst. and Stat., Lancaster, UK, 2002a.

Young, P. C., Advances in real-time flood forecasting, *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, *A360*, 1433–1450, 2002b.

K. Beven and P. Young, Environmental Science, Lancaster University, Lancaster LA1 4YQ, UK. (k.beven@lancaster.ac.uk)